Published online May 6, 2005

A Sampling Strategy for Conserving Genetic Diversity when Forming Core Subsets

Jorge Franco, José Crossa,* Suketoshi Taba, and Henry Shands

ABSTRACT

When forming core subsets, accessions from a collection are classified into clusters, and then samples are drawn from the clusters with the aim of maintaining the diversity of the collection. In a stratified sampling strategy, the allocation method provides a criterion for determining the number of accessions to be selected from each cluster. This paper proposes an allocation method (D method) and compares it with three other allocation methods (L, LD, and NY methods). In these allocation methods, the number of accessions sampled per cluster is proportional to (i) the mean of the Gower's distance between accessions within the cluster (D method), (ii) the logarithm of the cluster size (L method), (iii) the product of the cluster size times the mean Gower distance (NY method), and (iv) the product of the logarithm of the cluster size times the mean Gower distance (LD method). Five hundred independent stratified random samples with two sampling intensities (10 and 20%) were obtained from four datasets. The allocation methods were compared on the basis of three criteria: diversity of the samples, recovery of the range of variables in the sample, and variances of the samples. Results showed that the D method produced samples (i) with significantly more diversity than the other allocation methods, (ii) that recovered more of the range of the variables, (iii) with higher variances for the continuous variables than the other three methods, and (iv) with variances higher than the variance among accessions of the collection. A sampling intensity of 10% preserves the same or more variability than a sampling intensity of 20%.

Genetic resources stored in gene banks are usually sampled to foster efficient evaluation and utilization of the collections as well as to study phenotypic and genotypic diversity, form core subsets, and eliminate redundant and duplicate accessions within a collection. The main purpose of these activities is to preserve in the sample as much of the diversity present in the original collection as possible (Crossa et al., 1995a). For example, the approach of forming core collections (core subsets) was introduced to increase the efficiency of describing and using collections stored in gene banks, while preserving as much as possible the diversity of the entire collection (Frankel and Brown, 1984; Brown, 1989).

The process of sampling genetic resources with the objective of forming core subsets starts with grouping accessions to obtain homogeneous within and heterogeneous between clusters (or groups) and then using a predetermined sampling strategy within each cluster. The grouping of accessions into clusters is achieved by a classification strategy that partitions the original collection into groups with maximum distances between accessions located in different groups and minimum distances between accessions located in the same group. Franco et al. (1998, 1999, 2002) and Franco and Crossa (2002) proposed a sequential Ward-Modified Location Method (MLM) strategy in which the Gower (1971) distance is used as a measure of similarity (or distance) among accessions considering all continuous and categorical variables. The initial groups were formed by the Ward (1963) method, and then the MLM was used to improve those groups. The Ward-MLM strategy was used for analyzing the Latin American Maize Project (Taba et al., 1999) and Caribbean maize collections (Taba et al., 1998) with data from more than 10 countries and with the number of observations per collection ranging from 100 to 1800 and a mixture of continuous and discrete variables. These studies demonstrated that the Ward-MLM formed compact and well separated clusters.

The reason for sampling accessions when forming core subsets is to identify a strategy that will structure a sample that recovers most of the diversity contained in the original collection, while maximizing the variance and the distances between accessions in the sample. A sampling strategy involves defining a sampling intensity, a sampling method, and an allocation method (Thompson, 2002).

The sampling intensity defines the overall sample size, and for core collections, several authors studied sampling intensities that ranged from 5 to 20% of the total number of accessions (Brown, 1989; Schoen and Brown, 1993; Brown and Spillane, 1999; van Hintum, 1999; van Hintum et al., 2000). For species such as perennial ryegrass (*Lolium perenne* L.), Charmet and Balfourier (1995) found that a sampling intensity of 5 to 10% is optimal for capturing 86 to 90% of the diversity. However, for forming core collection of *Medicago* species, Diwan et al. (1995) pointed out that sampling intensities of 5 to 10% are insufficient to represent the original collection.

A stratified sampling method partitions the collection into clusters or groups, and then accessions within each cluster are selected. Several authors have recommended stratified sampling strategies for managing genetic resources and forming core subsets (Peeters and Martinelli, 1989; Crossa et al., 1994, 1995a; Spagnoletti Zeuli and Qualset, 1993; Charmet and Balfourier, 1995; Rincon et al., 1996). Statistical methods for stratifying genetic resources using three-way data (accessions × trait × location), with the purpose of forming core sub-

J. Franco, Facultad de Agronomía, Universidad de la República, Av. Garzón 780 CP 12900, Montevideo, Uruguay; J. Crossa, Biometrics and Statistics Unit, CIMMYT, Apdo. Postal 6-641, 06600, Mexico DF, Mexico; S. Taba, Maize Genetic Resources Unit, CIMMYT, Mexico; and H. Shands, National Center of Genetic Resources Preservation (NCGRP), USDA, ARS, Fort Collins, CO 80523. Received 12 May 2004. Plant Genetic Resources. *Corresponding author (j.crossa@cgiar.org).

Published in Crop Sci. 45:1035–1044 (2005).
doi:10.2135/cropsci2004.0292

677 S. Segoe Rd., Madison, WI 53711 USA

Abbreviations: DA, days to anthesis; DS, days to silking; EH, ear height; GM, grain moisture; MLM, Modified Location Model; PH, plant height.

sets, have been discussed by Crossa et al. (1995b) and, more recently, by Franco et al. (2003).

An allocation method provides criteria for determining the number of accessions to be selected from each cluster. For core subsets, Brown (1989) described three allocation methods whose sample sizes are (i) constant (or fixed) across clusters, (ii) proportional to the cluster size, and (iii) proportional to the logarithm of the cluster size. Brown (1989) also compared simple versus stratified sampling methods and recommended a stratified logarithmic method for choosing accessions from the collection. Finally, Brown (1989) proposed the logarithm of the cluster size (L method) as the allocation method. Yonezawa et al. (1995), Chandra et al. (2002), Diwan et al. (1995), and Zichao et al. (2002) have used the L method for sampling various crops. Diwan et al. (1994) formed core collections of 36 annual *Medicago* species and used an allocation method based on the diversity for the variables measured. The number of clusters formed in each species determined the diversity within species.

The main objectives of this study were to propose an allocation method (D method) for selecting accessions from the clusters (obtained by the Ward-MLM two stage strategy) and to compare it with other allocation methods (L, LD, and NY methods) with the aim of determining which one forms core subsets that best retain the diversity contained in the original collection. The four allocation methods determine sample size on the basis of different characteristics: (i) the D method: sample size proportional to the mean Gower distances between accessions within the cluster, (ii) the L method [proposed by Brown (1989)]: sample size proportional to the logarithm of the cluster size, (iii) the NY method [a modification of Neyman's (1934) method]: sample size proportional to the product of the cluster size times the mean Gower distance, and (iv) the LD method [a modification of Neyman's (1934) method]: sample size proportional to the product of the logarithm of the cluster size times the mean Gower distance. Five hundred independent stratified random samples under two sampling intensities, 10 and 20%, were obtained from three maize (*Zea mays* L.) collections and one maize population to compare the ability of the four allocation methods to retain the diversity of the collections.

MATERIALS AND METHODS

The Gower Distance

Gower (1971) proposed a similarity measure between the *i*th and the *j*th individuals, s_{ij}, that can use simultaneously continuous, ordinal, binary, and nominal variables. The author showed that a sufficient condition for the distance $[d_{ij} = (1 - s_{ij})^{1/2}]$ between two individuals to be a Euclidean metric is the positive semi-definite property of the similarity matrix $\mathbf{S} = \{s_{ij}\}$. In addition, the author showed that the similarity matrix **S** is positive semi-definite when there are no missing values in the data.

For k variables ($k = 1,2,...,p$), Gower's similarity measurement between two individuals i and j is:

$$s_{ij} = \frac{\sum_{k=1}^{p} w_{ijk} s_{ijk}}{\sum_{k=1}^{p} w_{ijk}}$$

where w_{ijk} is a weight given to the ijkth comparison, assigning values of 1 for valid comparisons and a value of 0 for invalid comparisons (when the value of the variable is missing in one or both individuals); s_{ijk} is the contribution of the kth variable to the total similarity between individuals i and j, and it takes values between 0 and 1. For a nominal variable, if the value of the kth variable is the same for both individuals, i and j, then $s_{ijk} = 1$; otherwise, it equals 0; for a continuous variable $s_{ijk} = 1 - |x_{ik} - x_{jk}|/R_k$ where x_{ik} and x_{jk} are the values of the kth variable for the i and j individuals, respectively, and R_k is the range (maximum value minus minimum value) of the kth variable in the sample. The division by R_k eliminates scale differences among variables, producing a value within the [0,1] interval and equal weights. The similarity value for binary characters is equal to the proportion of characters for which the two individuals agree, excluding the absence–absence agreement.

The Gower distance can be used as a diversity measure for a set of individuals (genotypes, accessions, etc.), with the important advantage that all types of variables can be used. Two genotypes with distances near zero show low diversity, whereas values near 1 indicate very diverse individuals.

The D Allocation Method

The D allocation method proposed in this study determines that the size of the sample to be drawn from each cluster should be proportional to the mean Gower distance between individuals within that cluster. Therefore, the number of accessions selected from each cluster will be proportional to the within-group diversity measured as the mean Gower distance between accessions within that group. More diverse groups will have a larger mean Gower's distance and therefore larger samples will have to be drawn from them.

For $t = 1,2,\ldots,g$ clusters, the number of accessions (n_t) to be drawn from the tth cluster (n_t) is

$$n_t = n \times p_t = n \times \frac{\overline{d}_t}{\sum_{t=1}^{g} \overline{d}_t} \qquad [1]$$

where n is the total sample size to be drawn from the collection (which in this study will be 10% or 20% of the entire collection), p_t is the proportion of the sample size to be drawn from the tth cluster, and $\overline{d}_t$ is the mean Gower distance between accessions within the tth cluster.

The L Allocation Method

The L allocation method uses the logarithm of the size of the cluster tth (N_t) to obtain the sample size of the tth cluster (n_t)

$$n_t = n \times \frac{log(N_t)}{\sum_{t=1}^{g} log(N_t)} \qquad [2]$$

with n as the total sample size (10 or 20%). The L method was proposed by Brown (1989) and later used by Yonezawa et al. (1995), Chandra et al. (2002), and Zichao et al. (2002).

The NY and LD Allocation Methods

Neyman (1934) proposed an optimal allocation method for estimating, with minimum variance, the mean value of the

variables in each cluster via stratified samples. The method determines that the size of the sample to be drawn from each cluster is proportional to the cluster size (N_t) and the standard deviation of the variable of interest, S_t, such that $n_t = n \times \frac{N_t S_t}{\sum_{t=1}^{g} N_t S_t}$. It recovers as much of the diversity present in the collection as possible by using the standard deviation of the variables in the cluster as the diversity measurement.

To make the Neyman (1934) optimal allocation method comparable with the other allocation methods, it was modified in two ways. First, the sample size of the tth cluster (N_t) was weighted by the diversity measured as the mean Gower distance ($\bar{d}_t$). This allocation method was named the NY method and is represented by

$$n_t = n \times \frac{N_t \bar{d}_t}{\sum_{t=1}^{g} N_t \bar{d}_t} \quad [3]$$

Second, to smooth out the effect of cluster size, the logarithm of N_t was weighted by the diversity of the tth cluster measured as the mean Gower distance ($\bar{d}_t$). This method was named the LD method

$$n_t = n \times \frac{log(N_t) \times \bar{d}_t}{\sum_{t=1}^{g} log(N_t) \times \bar{d}_t} \quad [4]$$

The Ward-MLM Sequential Clustering Strategy

The initial groups formed by any hierarchical (geometric) clustering technique are based on the principle that rules such a technique; for example, the minimum variance within groups of the initial technique is Ward. Geometric clustering methods can be used with continuous and/or discrete variables by means of Gower's distance.

Statistical classification methods use the concept of mixture models. An initial classification of the individuals into g clusters is given so that each group is one of the distributions in the mixture. The vector with the mean of the traits and the variance–covariance matrix within clusters are estimated by the maximum-likelihood method. The maximization of the likelihood function begins at a point that has been reached using the geometric technique; it will then reach a peak (which could be local) near the starting point that contains the characteristics of the geometric technique.

The Modified Location Model is a mixture model developed by Franco et al. (1998) that uses continuous and discrete variables simultaneously. The Ward-MLM sequential clustering strategy forms the initial groups using the Ward method and then improves them by the MLM, the idea being that the MLM method will modify the groups initially formed by the Ward method, so that the final classification is a statistical one.

The Ward strategy is the recommended geometric clustering method to use in the two-stage clustering strategy because (i) the objective function of the Ward strategy is to minimize the variance within clusters, whereas the objective function of the mixture distribution model is to maximize the likelihood of which the variance within a cluster is a component, (ii) the direct relationship between the Ward strategy and the multivariate analysis of variance technique are based on the result that the total variance is equal to the variance between clusters plus the variance within clusters, and (iii) the objective function of the Ward strategy allows producing spherical clusters, whereas the mixture distribution model allows the formation of clusters of another shape. Thus, the sequential clustering strategy allows the MLM to modify the form of the initial groups obtained by the Ward strategy to one that permits the formation of more homogeneous groups.

Determining the Number of Clusters in the Ward-MLM Method

The number of groups was determined by, first, the pseudo-F criterion (SAS Institute, 2000), which, for each division into g groups, the following ratio is computed:

$$\text{pseudo-F} = \frac{\text{tr}(\mathbf{B})/(g-1)}{\text{tr}(\mathbf{W})/(n-g)}$$

where tr(**B**) and tr(**W**) are the traces of the matrices of the sums of squares and cross products between and within groups, respectively. The number, g, of groups is selected in relation to the maximum value.

Then, we used the graph of the likelihood profile (related to the likelihood ratio test) for different values of g near the value obtained by the pseudo-F, and observed the maximum growth point of the likelihood profile as a criterion for determining the definitive number of groups. The optimal number of groups was then determined using the pseudo-F approach combined with the log-likelihood profile.

Datasets

In this study, three collections having different sizes (N), different values of diversity, and different numbers of clusters (g) were used (Taba et al., 1999). The Guatemalan collection had $N = 100$ accessions and the Ward-MLM strategy formed $g = 5$ clusters. The Brazilian collection comprised $N = 652$ accessions and the Ward-MLM strategy formed $g = 13$ clusters. The collection from Mexico had $N = 1460$ accessions and $g = 17$ were formed (Table 1). These datasets contained five continuous variables (days to anthesis, days to silking, plant and ear height, and grain moisture), two nominal variables (kernel color and texture) and two binary variables [number of ears per plant equals 0 when less than or equal to 1, and 1 when it was more than 1; ear quality rating (1–9) assigned the value of 0 when it was less than or equal to 4.5, and 1 when it was more than 4.5].

Another dataset, Pool 25 (Taba et al., 2001), with more variables than the other three, was also included ($N = 210$, $g = 7$) (Table 1). Pool 25 is a late tropical, yellow flint CIMMYT maize gene pool that comprises S2 lines crossed with a tester so that the entries should be very uniform. The 12 continuous variables were days to anthesis and silking, plant and ear height, days to senescence, grain moisture at harvest, shelling percentage, ear length and diameter, kernel row number by ear, and kernel length and width; the four binary variables were ear rot (0 = low, 1 = high), ear appearance (0 = bad, 1 = good), foliar disease score (0 = low, 1 = high), and agronomic scale (0 = bad, 1 = good).

Independent Stratified Random Samples

The allocation methods define how many, but not which specific, accessions per cluster should be sampled. The proposed D allocation method was evaluated and compared with

Table 1. Collection, number of accessions in the collection (N), number of clusters found by the Ward-MLM strategy (g), mean Gower distance between the N accessions of the entire collection ($\bar{d}$), mean Gower distance between accessions within clusters ($\bar{d}_t$).

Collection	N	g	$\bar{d}$	$\bar{d}_t$
Brazil	652	13	0.55	0.39
Guatemala	100	5	0.51	0.38
Mexico	1460	17	0.44	0.33
Pool 25	203	7	0.46	0.41

the L, LD, and NY allocation methods by randomly drawing 500 samples from three maize collections and one maize gene pool. First, accessions from each of the four datasets were classified by Ward-MLM. Second, from each classified dataset, 500 independent stratified random samples (without replacement) were drawn, for each of the factorial combinations of two sampling intensities (10 and 20% of the entire collection) and the four allocation methods (D, L, LD, and NY). This was done by the SURVEYSELECT procedure of SAS (SAS Institute, 2000) and a computational code written in SAS procedure in IML (SAS Institute, 2000). Values were computed for the criteria used to compare the four allocation methods (see below). For each of the 500 samples, accessions within each cluster in each of the four datasets were selected at random.

Criteria for Comparing the Allocation Methods

A sampling strategy aims (i) to define a sampling intensity and an allocation method that will retain in the sample most of the collection diversity and (ii) to produce a sample with maximum variance and maximum distance between accessions, as compared with the variance and distances between accessions in the entire collection. The criteria we used for comparing the D method with the L, LD, and NY methods are described as follows.

Diversity of the Sample

The best allocation method is the one that produces a sample with a greater mean Gower distance among accessions ($\overline{d}_S$). For allocation methods, sampling intensities, and allocation method × sampling intensity interactions, the mean Gower distances across 500 independent random samples were statistically compared.

Recovery of the Range in the Sample

The recovery of the range (RR) for all variables (discrete and continuous) is given by $RR = \frac{1}{p}\sum_{k=1}^{p}\frac{Rn_k}{RN_k}$, where Rn_k, and RN_k are the ranges of the kth variable in the sample and in the entire collection, respectively, for $k = 1, 2, \ldots, p$ variables. An allocation method is better if it selects a sample with an RR near 1. The mean recovery of the range ($\overline{RR_s}$) values for allocation methods, sampling intensities, and allocation method × sampling intensity interactions were also statistically compared.

Variances of the Samples

An optimal allocation method should produce samples with high variance among the accessions. The variance of the accessions in the sample was measured for the five continuous variables: days to anthesis (DA), days to silking (DS), plant height (PH), ear height (EH), and grain moisture (GM). Thus, differences in the mean variances of each continuous variable, $\overline{S^2_{DA}}, \overline{S^2_{DS}}, \overline{S^2_{PH}}, \overline{S^2_{EH}}, \overline{S^2_{GM}}$, for allocation methods, sampling intensities, and allocation method × sampling intensity interactions were statistically assessed.

Comparing Allocation Methods

Analyses of variance for each dataset considered the allocation method, the sampling intensity, and the allocation method × sampling intensity interaction as fixed effects. Comparisons between allocation methods were performed across sample intensities and within sampling intensity. The dependent variables were the criteria used to evaluate the allocation methods: diversity of the sample measured by the mean Gower distance among accessions in the sample ($\overline{d}_s$), the recovery of the range in the sample ($\overline{RR_s}$), and the variance of the sample for five continuous variables DA, DS, PH, EH, and GM ($\overline{S^2_{DA}}, \overline{S^2_{DS}}, \overline{S^2_{PH}}, \overline{S^2_{EH}}, \overline{S^2_{GM}}$, respectively).

Pairwise comparisons of allocation methods across sampling intensities and within sampling intensity were made for $\overline{d}_s$, $\overline{RR_s}$, $\overline{S^2_{DA}}$, $\overline{S^2_{DS}}$, $\overline{S^2_{PH}}$, $\overline{S^2_{EH}}$, and $\overline{S^2_{GM}}$ using the Tukey's studentized range test.

Ranking the Allocation Methods

The Friedman two-way test (Conover, 1971) was performed, within each sample intensity, for testing the null hypothesis

H_o: each ranking within the seven response variables: $\overline{d}_s$, $\overline{RR_s}$, $\overline{S^2_{DA}}$, $\overline{S^2_{DS}}$, $\overline{S^2_{PH}}$, $\overline{S^2_{EH}}$, and $\overline{RR_s}$ is equally likely (i.e., there is not a consistent order among allocation methods) versus the alternative hypothesis,

H_a: at least one of the allocation methods tended to perform consistently better (i.e., there are a consistent order among allocation methods).

Comparing Allocation Methods with the Entire Collection

On the basis of the criteria described above, we compared the four allocation methods with the entire collection in each of the 500 independent stratified random samples.

It is expected that the mean Gower distance between accessions in the sample is greater than that between accessions in the entire collection. This is due to the fact that while the sample preserves diversity, it also has fewer redundant accessions. Thus, if the sample has a good representation of the diversity in the collection but fewer redundant accessions, its mean Gower distance will be greater than the mean Gower distance in the entire collection. If the mean Gower distance between accessions of the entire collection is $\overline{d}_c$, then a good performance criterion is when the mean Gower distance between the selected accessions forming the sample ($\overline{d}_s$) is greater than $\overline{d}_c + 0.1\overline{d}_c$ or $\overline{d}_c + 0.2\overline{d}_c$ or $\overline{d}_c + 0.3\overline{d}_c$.

Concerning the recovery of the range (RR) of the variables in the sample, an allocation method is better if it selects a sample with high RR. Regarding the variances of the variables in the sample, a procedure is better if it produces samples with higher variances than the variance among accessions in the entire collection. We used the criteria $S^2_S \geq [S^2_C + 0.1S^2_C]$, $S^2_S \geq [S^2_C + 0.2S^2_C]$, and $S^2_S \geq [S^2_C + 0.5S^2_C]$ where S^2_S and S^2_C are the variances for the sample and the entire collection, respectively, for each continuous variable. In the sampling study, the number of times that $S^2_S \geq [S^2_C + 0.1S^2_C]$, $S^2_S \geq [S^2_C + 0.2S^2_C]$, and $S^2_S \geq [S^2_C + 0.5S^2_C]$ were recorded.

RESULTS AND DISCUSSION

The Ward-MLM method produced clusters with smaller mean Gower distances ($\overline{d}_t$) between accessions within each cluster than the average of the Gower distances between accessions in the entire collection ($\overline{d}$) for the four datasets (Table 1). The dataset from Mexico showed the highest number of observations, number of groups, and the lowest values for within cluster ($\overline{d}_t$ = 0.33) and total average ($\overline{d}$ = 0.44) distances. The Guatemala dataset had the lowest number of observations and lowest number of groups, whereas the Brazil and

Pool 25 datasets had the highest values for $\bar{d}$ and $\bar{d}_t$, respectively. The values of $\bar{d}_t$ for each individual cluster in all datasets were always smaller than the average distance between accessions in the entire collection ($\bar{d}$), except for two clusters (3 and 5) in the Mexico collection (Table 2). When the allocation method requires a sample size larger than the size of the cluster then fewer accessions will be sampled. This is the case in the Mexico collection where the D method resulted in selecting fewer accessions from cluster 5 (17) than clusters 2, 3, 9, 10, 15, and 17, even though cluster 5 had the greatest $\bar{d}_t$. These results indicate that the Ward-MLM sequential clustering strategy formed homogeneous groups.

Table 2 shows that, for Groups 4 and 5 from Mexico, the D and LD methods required a sample size equal to or larger than the group size because of the heterogeneity of the groups (high distance values) combined with a small group size. In these cases, the entire cluster was included. In Pool 25, the D method allocated the same number of accessions to all groups, and the mean Gower distances within clusters were very similar, ranging from 0.37 to 0.42 (Table 2). For Pool 25, the other methods did not allocate a similar number of accessions per cluster, as did the D method. These results are in agreement with the high uniformity of the entries comprising Pool 25.

In general, the NY method tends to form groups of very different sizes. For example, in the Mexico collection the group size ranged from 3 to 73. In contrast, methods D and LD formed groups less diverse in size. For example, with the D method, the size of the groups ranged from 13 to 24, and with LD method, from 13 to 25.

The size of samples drawn from each cluster using the D allocation method is based on the diversity of the cluster ($\bar{d}_t$) and not on its size (N_t) (Table 2). For example, for the Mexico collection, Group 6 had $N_t = 450$ accessions with the lowest diversity $\bar{d}_t = 0.25$; the D method allocated 13 accessions to this group, whereas LD, NY, and L methods allocated 21, 73, and 28 accessions, respectively. On the other hand, Mexico Groups 3 and 5 had $N_t = 29$ and $N_t = 17$ accessions, respectively, and the two highest diversity values: $\bar{d}_t = 0.47$ and $\bar{d}_t = 0.48$, respectively; the D method allocated 24 and 17 accessions to Groups 3 and 5, respectively, but the other allocation methods assigned a smaller number of accessions to these clusters. Similarly, for the Brazil collection, Group 9 had $N_t = 106$ and $\bar{d}_t = 0.24$ and Group 13 comprised $N_t = 50$ and had $\bar{d}_t = 0.48$; the D method assigned 6 accessions to Group 9 and 12 to Group 13.

Comparing Allocation Methods

Diversity of the Sample

The mean Gower distances between accessions across the 500 samples ($\bar{d}_s$) were higher than the respective mean Gower distance between accessions in the entire collection for the four datasets and for each allocation method–sampling intensity combination (Table 3). The minimum value of the 500 samples for all datasets and allocation methods was always larger than the mean Gower distance between accessions of the corresponding datasets. These results indicate that all allocation methods selected samples formed by a well-differentiated group of accessions.

The analysis of variance showed that there were significant differences ($P \leq 0.01$) between levels of allocation method, sampling intensity, allocation method × sampling intensity interaction and allocation methods within sampling intensities effects (data not shown). For all datasets and both sampling intensities, the Tukey's test indicated that $\bar{d}_s$ of the D method was always significantly higher ($P \leq 0.01$) than $\bar{d}_s$ of the other allocation methods (Table 3). When combining the allocation methods across both sampling intensities, $\bar{d}_s$ of the D method was significantly superior to $\bar{d}_s$ of the other allocation methods for all datasets (data not shown). For all data-

Table 2. Sample size (n_t) for the four allocation methods (D, LD, NY, and L) for a 20% sampling intensity and four datasets. Number of clusters (g), number of observations per cluster (N_t), and mean Gower distance per cluster ($\bar{d}_t$).

	Mexico						Brazil						Pool 25						Guatemala					
g	N_t	$\bar{d}_t$	D	LD	NY	L	N_t	$\bar{d}_t$	D	LD	NY	L	N_t	$\bar{d}_t$	D	LD	NY	L	N_t	$\bar{d}_t$	D	LD	NY	L
												n_t												
1	52	0.25	13	14	9	18	20	0.40	10	8	4	8	32	0.42	6	6	7	6	29	0.33	3	4	5	5
2	37	0.44	23	22	11	17	69	0.43	11	13	16	11	32	0.41	6	6	6	6	40	0.41	4	6	9	5
3	29	0.47	24	22	9	16	30	0.45	12	11	7	9	34	0.40	6	6	7	6	4	0.25	3	1	1	2
4	14	0.39	14	14	4	12	30	0.42	11	10	7	9	49	0.41	6	7	10	7	10	0.48	5	4	2	3
5	17	0.48	17	17	5	13	71	0.34	9	10	14	11	25	0.42	6	6	5	6	17	0.40	4	4	4	4
6	450	0.25	13	21	73	28	39	0.45	12	12	10	10	31	0.41	6	6	6	6	–	–	–	–	–	–
7	76	0.29	15	17	14	20	13	0.37	9	7	3	7	7	0.37	6	3	1	4	–	–	–	–	–	–
8	1	–	1	1	1	1	77	0.26	7	8	11	12	–	–	–	–	–	–	–	–	–	–	–	–
9	41	0.36	19	18	10	17	106	0.24	6	8	14	12	–	–	–	–	–	–	–	–	–	–	–	–
10	120	0.34	18	22	27	22	73	0.28	7	8	11	11	–	–	–	–	–	–	–	–	–	–	–	–
11	85	0.23	12	14	13	21	42	0.48	12	12	11	10	–	–	–	–	–	–	–	–	–	–	–	–
12	13	0.38	13	13	3	12	32	0.46	12	11	8	9	–	–	–	–	–	–	–	–	–	–	–	–
13	254	0.32	17	25	54	26	50	0.48	12	13	13	10	–	–	–	–	–	–	–	–	–	–	–	–
14	199	0.33	17	24	43	25	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
15	21	0.38	20	16	5	14	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
16	21	0.33	17	14	5	14	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
17	30	0.39	20	18	8	16	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	1460†	0.44‡	273	292	293	292	652†	0.55‡	130	131	129	129	210†	0.46‡	42	40	42	41	100†	0.51‡	19	19	21	19

† Number of accessions in each collection (N in Table 1).
‡ Mean Gower distance of the N accessions of the collection ($\bar{d}$ in Table 1).

Table 3. Mean Gower distance between the accessions of the sample ($\bar{d}_s$), mean recovery of the range in the sample ($\overline{RR_s}$), mean of the variance for days to anthesis ($\overline{S^2_{DA}}$), days to silking ($\overline{S^2_{DS}}$), plant height ($\overline{S^2_{PH}}$), ear height ($\overline{S^2_{EH}}$) and grain moisture ($\overline{S^2_{GM}}$) for two sampling intensities (10 and 20%), four allocation methods (D, LD, L, and NY) for four datasets and for the entire collection (Coll.). Mean rank across $\bar{d}_s$, $\overline{RR_s}$, $\overline{S^2_{DA}}$, $\overline{S^2_{DS}}$, $\overline{S^2_{PH}}$, $\overline{S^2_{EH}}$, $\overline{S^2_{GM}}$ ($\overline{Rank}$), and chi-square value for the Friedman test (χ^2).

Int.	Met.	$\bar{d}_s$	$\overline{RR_s}$	$\overline{S^2_{DA}}$	$\overline{S^2_{DS}}$	$\overline{S^2_{PH}}$	$\overline{S^2_{EH}}$	$\overline{S^2_{GM}}$	$\overline{Rank}$	χ^2
					Guatemala					
10	D	0.598a†	0.859a	194.2a	217a	3574a	1597a	17.0a	1.43	
10	LD	0.586b	0.836b	166.3b	173b	3176b	1489b	17.3a	2.29	
10	L	0.579c	0.847c	156.6c	161c	3004c	1445b	17.7a	2.86	
10	NY	0.556d	0.731d	149.9c	155c	2928c	1465b	17.8a	3.43	9.2*
20	D	0.574a	0.916a	181.4a	199a	3395a	1545a	17.9a	1.29	
20	LD	0.560b	0.894b	134.4b	128b	2662b	1311b	18.1a	2.86	
20	L	0.558b	0.910ab	160.9c	169c	3020c	1449c	17.9a	2.14	
20	NY	0.537c	0.901b	131.3b	124b	2561b	1338b	17.7a	3.71	13.5**
Coll.		0.506	1.000	123.2	114	2409	1282	17.6		
					Brazil					
10	D	0.610a	0.890a	48.3a	45.1a	1719a	755a	0.158a	1.00	
10	LD	0.608b	0.886a	46.9b	43.5b	1632b	735b	0.154ac	2.57	
10	L	0.603c	0.885a	46.4b	44.2b	1608b	747ab	0.156ab	2.71	
10	NY	0.597d	0.736b	45.2c	42.3c	1518c	750a	0.152c	3.71	15.9**
20	D	0.601a	0.925a	47.9a	44.6a	1724a	749a	0.157a	1.14	
20	LD	0.598b	0.925a	47.0ab	43.8a	1652b	745a	0.154a	2.43	
20	L	0.593c	0.921ab	46.5b	44.2a	1620c	751a	0.156a	2.43	
20	NY	0.585d	0.918b	44.5c	41.7b	1489d	743a	0.148b	4.00	17.2**
Coll.		0.539	1.000	43.2	41.5	1355	700	0.147		
					Mexico					
10	D	0.549a	0.969a	454.5a	445a	2266a	1800a	57.5a	1.00	
10	LD	0.538b	0.963b	404.3b	396b	2108b	1651b	55.1b	2.00	
10	L	0.526c	0.960c	359.3c	353c	1957c	1533c	52.9c	3.00	
10	NY	0.473d	0.943d	190.9d	190d	1565d	1150d	52.1d	4.00	21.0**
20	D	0.545a	0.988a	455.4a	445a	2293a	1804a	57.8a	1.00	
20	LD	0.533b	0.985b	399.4b	391b	2096b	1643b	55.4b	2.00	
20	L	0.522c	0.981c	363.3c	356c	1976c	1539c	53.0c	3.00	
20	NY	0.464d	0.961d	171.4d	171d	1529	1118d	51.3d	4.00	21.0**
Coll.		0.440	1.000	152.7	152.8	1466	1065	48.9		
					Pool-25					
10	D	0.540a	0.467a	3.999a	4.02a	93.4a	89.8a	5.00a	1.00	
10	LD	0.536b	0.461b	3.492b	3.56b	92.5a	91.8ab	4.80a	2.00	
10	L	0.536c	0.461b	3.492b	3.56b	92.5a	91.8ab	4.80a	2.00	
10	NY	0.533d	0.457b	2.863c	2.99c	90.7a	94.8b	4.80a	4.00	21.0**
20	D	0.512a	0.506a	3.957a	3.99a	92.9a	90.2a	5.10a	1.43	
20	LD	0.508b	0.494b	3.194b	3.28b	91.4a	92.6a	4.83b	3.00	
20	L	0.508b	0.500c	3.484c	3.56c	92.3a	91.5a	4.91ab	2.29	
20	NY	0.510c	0.483d	2.550a	2.71d	90.3a	95.7a	4.66c	3.29	8.7*
Coll.		0.464	0.563	2.692	2.84	88.3	93.3	4.70		

*** Mean ranks were consistent and significantly different by the Friedman test at $P \leq 0.05$.**
**** Mean ranks were consistent and significantly different by the Friedman test at $P \leq 0.01$.**
† Means with different letters within each sampling intensity are significantly different by the Tukey's test at $P \leq 0.01$.

sets, the $\bar{d}_s$ of the D method produced with sampling intensity of 10% was significantly higher than the $\bar{d}_s$ of samples generated with 20% sampling intensity (data not shown).

The distribution of the mean Gower distances (mean D) from 500 samples is shown as box plots in Fig. 1. The D method produced the highest values for all datasets and for both sampling intensities (10% and 20%). In general, a 10% sampling intensity generated samples with higher mean Gower distance than the 20% sampling intensity, for all allocation methods and collections. Thus, for these datasets and this diversity criterion, a 20% sampling intensity resulted in redundant information, and the 10% sampling intensity was sufficient for representing collection's diversity.

Recovery of the Range in the Sample

There were significant differences ($P \leq 0.01$) between levels of allocation method, sampling intensity, allocation method × sampling intensity interaction and allocation methods within sampling intensities effects in all datasets (data not shown). The Tukey's test indicated that $\overline{RR_s}$ of the D method was always significantly higher ($P \leq 0.01$) than $\overline{RR_s}$ of the other allocation methods (Table 3) in all datasets except Brazil in both sampling intensities. Averaged across sampling intensities, the D method had $\overline{RR_s}$ values significantly larger than the $\overline{RR_s}$ values of the other allocation methods for all datasets except Brazil ($\overline{RR_s}$ of the D and L methods were similar). In all datasets, the $\overline{RR_s}$ for 20% sampling intensity (across allocation methods) was significantly larger than the $\overline{RR_s}$ for 10% sampling intensity (data not shown).

The distribution of the *RR* values from 500 samples is shown as box plots in Fig. 2. In general, a 20% sampling intensity generated samples with better *RR* values than the 10% sampling intensity, for all allocation methods and collections (Fig. 2).

Variances of the Samples

Not all the effects (sampling intensity, allocation method × sampling intensity interaction and allocation

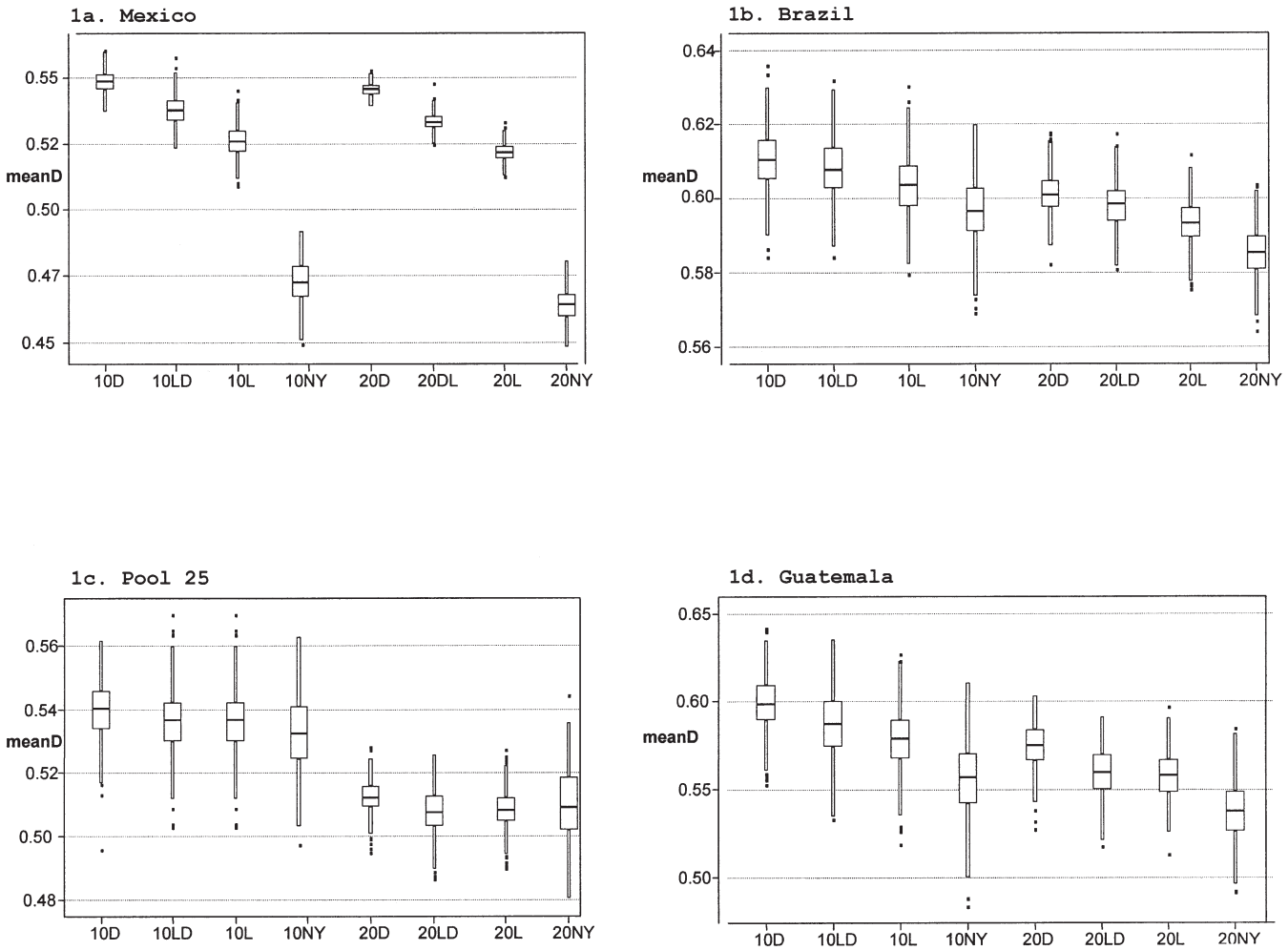

Fig. 1. Box plot representation of the mean Gower distance among accessions (meanD) for 500 samples using four allocation methods (D, LD, L, and NY) and two sampling intensities: 10%, and 20% (10D, 10LD, 10L, 10NY, 20D, 20LD, 20L, 20NY) for Mexico (1a), Brazil (1b), Pool 25 (1c), and Guatemala (1d) collections.

methods nested within sampling intensities effects) were significantly different for all mean variances of the five continuous variables in all the datasets. Only the different allocation methods were significantly different ($P \leq 0.01$) for all datasets for the mean variances of the five variables. The Tukey's test indicated that the values of $\overline{S^2_{DA}}$, $\overline{S^2_{DS}}$, $\overline{S^2_{PH}}$, $\overline{S^2_{EH}}$, and $\overline{S^2_{GM}}$ were significantly larger with the D method than the other methods in most cases, except for: 1) $\overline{S^2_{GM}}$ in Guatemala, Brazil, and Pool 25 for both sampling intensities; 2) $\overline{S^2_{DA}}$, $\overline{S^2_{DS}}$, and $\overline{S^2_{EH}}$ in Brazil for 20% sampling intensity; 3) $\overline{S^2_{PH}}$ and $\overline{S^2_{EH}}$ in Pool 25 for 10% and 20% sampling intensities (Table 3).

The mean variances of the variables for all datasets and allocation methods tended to be larger for 10% sampling intensity than for 20% sampling intensity (Table 3). When the allocation methods are averaged across sampling intensities, the values of $\overline{S^2_{DA}}$ and $\overline{S^2_{DS}}$ for the D method were significantly larger than those of the other allocation methods (data not shown). For $\overline{S^2_{PH}}$ and $\overline{S^2_{EH}}$ the D method significantly differed from the other methods, except in Pool 25. For $\overline{S^2_{GM}}$ the D method differed from the others only in Mexico and Pool 25.

Ranking the Allocation Methods

The D allocation method ranked consistently first for $\overline{d}_s$ and $\overline{RR}_s$ variables for all datasets and sample intensities. The D method ranked first in most of the variances of the five continuous variables, except for $\overline{S^2_{GM}}$ in Guatemala under both sample intensities and for $\overline{S^2_{EH}}$ in Brazil and Pool 25 under 20% sample intensity. The mean rank of each allocation method in each dataset and sample intensity is shown in Table 3. The Friedman test for each dataset and sample intensity determined that the data are consistent with the hypothesis that the D allocation method performed consistently higher than the other allocation methods for all seven variables.

Comparing Allocation Methods with the Entire Collection

Diversity of the Sample

Across datasets and sampling intensities, the D allocation method produced a larger percentage of samples with $\overline{d}_s \geq [\overline{d}_c + 0.1\overline{d}_c]$ than the other allocation methods at both sampling intensities (Table 4). For the interval

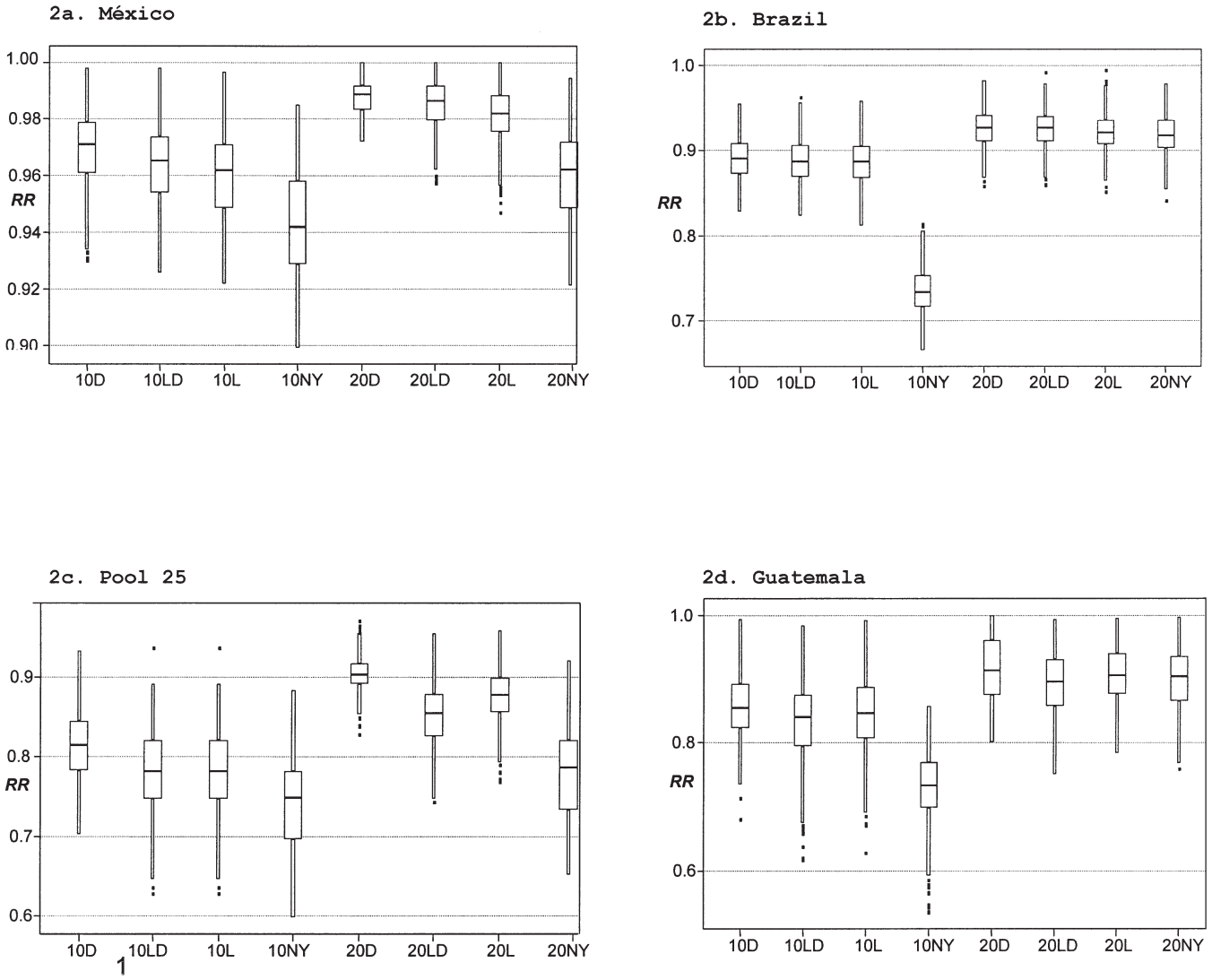

Fig. 2. Box plot representation of the recovery of the range (*RR*) for 500 samples using four allocation methods (D, LD, L, and NY) and two sampling intensities: 10%, and 20% (10D, 10LD, 10L, 10NY, 20D, 20LD, 20L, 20NY) for Mexico (2a), Brazil (2b), Pool 25 (2c), and Guatemala (2d) collections.

$[\bar{d}_c + 0.2\bar{d}_c]$ the D method was superior to the other methods only in Mexico (at both sampling intensities) and in Guatemala with 10% sampling intensity.

Recovery of the Range in the Sample

The D method produced the same or a higher number of samples that recovered 80% (RR_{80}) and 90% (RR_{90}) of the range of variables included in the analysis than were produced by the other allocation methods, for all datasets and sampling intensities (Table 4). The exception was the Brazil collection with 10% sampling intensity, where the D method recovered 90% of the range in only 34% of the 500 samples, as compared with the NY method, which recovered 90% of the range in all 500 samples (Table 4).

Variances of the Samples

For all datasets and sampling intensities, the D method resulted in the highest percentage of the 500 samples with variances among the accessions in the sample (S_S^2) that were greater than the values for $[S_C^2 + 0.1S_C^2]$, $[S_C^2 + 0.2S_C^2]$, and $[S_C^2 + 0.5S_C^2]$ (data not shown). The only exception was for the variable GM in the Guatemala collection. It is interesting that for all datasets, the D method tended to generate more diverse samples than the other methods as the width of the interval increased from 10% to 50%. These results indicate that the D method produced samples with maximum variance and maximum distance between accessions as compared with the variance and the distances between accessions in the entire collection.

CONCLUSIONS

This research proposes the D allocation method and compares it with other allocation methods with the objective of forming core subsets that will capture and, therefore, represent most of the diversity existing in the original collection. The D allocation method seems to

Table 4. Percentage of the 500 samples showing a mean Gower distance between accessions ($\bar{d}_s$) greater than [$\bar{d}_c + 0.1\bar{d}_c$], and [$\bar{d}_c + 0.2\bar{d}_c$] ($\bar{d}_c$ = mean Gower distance between accessions of the entire collection) for two sampling intensities (10% and 20%), four allocation methods (D, LD, L, and NY) and four data sets. Percentage of samples showing a Recovery of the Range (*RR*) greater than 0.80(RR_{80})and 0.90(RR_{90}).

Sampling intensity	Allocation method	RR_{90}	RR_{80}	[$\bar{d}_c + 0.1\bar{d}_c$]	[$\bar{d}_c + 0.2\bar{d}_c$]†
		%			
		Mexico			
10	D	100	100	100	100
10	LD	100	100	100	98
10	L	100	100	100	37
10	NY	100	100	9	0
20	D	100	100	100	100
20	LD	100	100	100	98
20	L	100	100	100	3
20	NY	100	100	0	0
		Brazil			
10	D	34	100	99	0
10	LD	30	100	99	0
10	L	31	100	90	0
10	NY	100	100	69	0
20	D	88	100	94	0
20	LD	86	100	81	0
20	L	84	100	53	0
20	NY	78	100	13	0
		Pool 25			
10	D	2	62	100	1
10	LD	1	39	99	3
10	L	1	39	99	3
10	NY	0	18	98	4
20	D	58	100	65	0
20	LD	12	89	36	0
20	L	25	98	34	0
20	NY	1	39	46	0
		Guatemala			
10	D	20	87	100	29
10	LD	14	73	91	15
10	L	17	79	90	5
10	NY	0	10	50	0
20	D	57	100	90	0
20	LD	46	98	58	0
20	L	55	100	53	0
20	NY	54	98	11	0

† For [$\bar{d}_c + 0.3\bar{d}_c$] all percentages are zero.

be effective in structuring samples that will preserve the diversity of the original collection. In the three collections and Pool 25 and with both sampling intensities, the D method resulted in significantly larger mean Gower distances between accessions in the samples than the mean Gower distances between accessions in the samples obtained with other allocation methods. Results indicated that the D allocation method recovered significantly more of the range of variables in the sample than did the other allocation methods. In general, the D method generated samples with significantly larger variance than the other methods. The exception was grain moisture.

In most cases, for the response variables $\bar{d}_s$, $\overline{RR_S}$, $\overline{S^2_{DA}}$, $\overline{S^2_{DS}}$, $\overline{S^2_{PH}}$, $\overline{S^2_{EH}}$, and $\overline{S^2_{GH}}$, the D allocation method ranked first. The mean rank of the D allocation method was statistically higher than the mean rank of the other allocation methods. Concerning the sampling intensities, the results of this study indicated that for $\bar{d}_s$, $\overline{S^2_{DA}}$, $\overline{S^2_{DS}}$, $\overline{S^2_{PH}}$, $\overline{S^2_{EH}}$, $\overline{S^2_{GM}}$ a sample of 10% of the entire collection is sufficient for preserving the diversity of the collection, whereas results based on $\overline{RR_S}$ showed that a sampling intensity of 20% preserves more of the diversity.

In this study, accessions from each cluster were randomly selected according to the sample size determined by the four allocation methods. However, allocation methods do not define which specific accessions should be sampled. Accessions can be selected from each cluster on the basis of other criteria as well, such as general agronomic performance, grain yield, and general plant type. Some researchers may decide to select the best performing accessions to be crossed with line testers or elite germplasm sources, and then initiate a prebreeding program. For example, the D method can be combined with an agronomic selection criterion for selecting accessions from each cluster.

The D method can be used with any clustering strategy and any distance measure. In this study the clustering strategy was the Ward-MLM used with continuous and discrete variables; the only distance that can be used for such data is Gower's distance, which is thus the distance that should be used in the D allocation method. The D method may be useful not only for sampling genetic diversity in crop germplasm collections but also in other areas of research where a stratified sampling method is required for preserving as much of the original population's diversity as possible.

The Ward-MLM strategy can use phenotypic and genetic marker data simultaneously, as shown by Franco et al. (2001). Using only molecular markers and/or DNA sequence data, various genetic distances and hierarchical clustering algorithms can be employed, and various allocation methods evaluated. Results can be validated based on phenotypic data, as was done by McKhann et al. (2004). However, further research is needed to assess the usefulness of the D allocation method using only marker data and to compare it with other allocation methods that do not use stratified sampling such as the M (maximization) strategy proposed by Schoen and Brown (1993).

REFERENCES

Brown, A.H.D. 1989. Core collections: A practical approach to genetic resources management. Genome 31:818–824.

Brown, A.H.D., and C. Spillane. 1999. Implementing core collections principles procedures, progress, problems and promise. p. 1–9. *In* R.C. Johnson and T. Hodgkin (ed.) Core collections for today and tomorrow. International Plant Genetic Resources Institute, Rome.

Chandra, S., Z. Huaman, Z., S. Harish Krishna, and R. Ortiz. 2002. Optimal sampling strategy and core collection size of Andean tetraploid potato based on isozyme data–A simulation study. Theor. Appl. Genet. 104:1325–1334.

Charmet, G., and F. Balfourier. 1995. The use of geostatistics for sampling a core collection of perennial ryegrass population. Genet. Resour. Crop Evol. 42:303–309.

Conover, W.J. 1971. Practical nonparametric statistics. Wiley & Sons Inc., New York.

Crossa, J., S. Taba, S.A. Eberhart, P. Bretting, and R. Vencovsky. 1994. Practical considerations for maintaining germplasm in maize. Theor. Appl. Genet. 89:89–95.

Crossa, J., I.H. DeLacy, and S. Taba. 1995a. The use of multivariate methods in developing a core collection. p. 77–89. *In* T. Hodgkin et al. (ed.) Core collections of Plant genetic resources. John Wiley & Sons Inc., New York.

Crossa, J., K. Basford, S. Taba, I. DeLacy, and E. Silva. 1995b. Three-

Mode analysis of maize using morphological and agronomic attributes measured in multilocation Trials. Crop Sci. 35:1483–1491.
Diwan, N., G.R. Bauchan, and M. Mcintosh. 1994. A core collection for the United States annual *Medicago* germplasm collection. Crop Sci. 34:279–285.
Diwan, N., M.S. McIntosh, and G.R. Bauchan. 1995. Methods of developing a core collection of annual *Medicago* species. Theor. Appl. Genet. 90:755–761.
Franco, J., J. Crossa, J. Villaseñor, S. Taba, and S.A. Eberhart. 1998. Classifying genetic resources by categorical and continuous variables. Crop Sci. 38:1688–1696.
Franco, J., J. Crossa, J. Villaseñor, S. Taba, and S.A. Eberhart. 1999. A two-stage, three-way method for classifying genetic resources in multiple environments. Crop Sci. 39:259–267.
Franco, J., J. Crossa, J.M. Ribaut, J. Betran, M.L. Warburton, and M. Khairallah. 2001. A method for combining molecular markers and pehotypic attributes for classifying plant genotypes. Theor. Appl. Genet. 103:944–952.
Franco, J., and J. Crossa. 2002. The Modified Location Model for classifying genetic resources. I: Association between categorical and continuous variables. Crop Sci. 42:1719–1726.
Franco, J., J. Crossa, S. Taba, and S.A. Eberhart. 2002. The Modified Location Model for classifying genetic resources. II: Unrestrictive variance-covariance matrices. Crop Sci. 42:1727–1736.
Franco, J., J. Crossa, S. Taba, and H. Shands. 2003. A multivariate method for classifying cultivars and studying group × environment × trait interaction. Crop Sci. 43:1249–1258.
Frankel, O.H., and A.H.D. Brown. 1984. Plant genetic resources today: A critical appraisal. p. 249–257. *In* J.H.W. Holden and J.T. Williams (ed.) Crop genetic resources: Conservation and evaluation. George Allen and Unwin London.
Gower, J.C. 1971. A general coefficient of similarity and some of its properties. Biometrics 27:857–874.
McKhann, H.I., C. Camilleri, A. Berárd, T. Bataillon, J.L. David, X. Reboud, V. Le Corre, C. Caloustain, I.G. Gut, and D. Brunel. 2004. Nested core collections maximizing genetic diversity in *Arabidopsis thaliana.* Plant J. 38:193–202.
Neyman, J. 1934. On the two different aspects on the representative method: The method of stratified sampling and the method of purposive selection. J. R. Statist. Soc. 97:558–606.
Peeters, J.P., and J.A. Martinelli. 1989. Hierarchical cluster Analysis as a tool to manage variation in germplasm collections. Theor. Appl. Genet. 78:42–48.
Rincon, F., B. Johnson, J. Crossa, and S. Taba. 1996. Cluster analysis, and approach to sampling variability in maize accessions. Maydica 41:307–316.
SAS Institute Inc. 2000. SAS/STAT, Version 8.2, Cary, NC.
Schoen, D.J., and A.H.D. Brown. 1993. Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. Proc. Natl. Acad. Sci. USA 90:10623–10627.
Spagnoletti Zeuli, P.L., and C.O. Qualset. 1993. Evaluation of five strategies for obtaining a core subset from a large genetic resource collection of durum wheat. Theor. Appl. Genet. 87:295–304.
Taba, S., J. Díaz, J. Franco, and J. Crossa. 1998. Evaluation of Caribbean maize accessions to develop a core subset. Crop Sci. 38:1378–1386.
Taba, S., J. Díaz, J. Franco, J. Crossa, and S.E. Eberhart. 1999. A Core subset of LAMP, from the Latin American Maize Project. CIMMYT, Mexico DF.
Taba, S., K. Duncan, M. Krakowsky, and J. Diaz. 2001. Annual research report, maize genetic resources. CIMMYT, México DF.
Thompson, S.K. 2002. Sampling. 2nd ed. John Wiley & Sons, New York.
van Hintum, Th. J.L., A.H.D. Brown, C. Spillane, and T. Hodgkin. 2000. Core Collections of Plant genetic resources. IPGRI Tech. Bull. 3. International Plant Genetic Resources Institute, Rome.
van Hintum, Th. J.L. 1999. The general methodology for creating a core collection. p. 10–17. *In* R.C. Johnson and T. Hodgkin (ed.) Core collections for today and tomorrow. International Plant Genetic Resources Institute, Rome.
Ward, J.H., Jr. 1963. Hierarchical grouping to optimize an objective function. J. Am. Statist. Assoc. 58:236–244.
Yonezawa, K., T. Nomura, and H. Morishima. 1995. Sampling strategies for use in stratified germplasm collections. p. 35–53. *In* T. Hodgkin et al (ed.) Core collections of plant genetic resources. Jon Wiley & Sons, Inc., New York.
Zichao, L., Z. Hongliang, Z. Yawen, Y. Zhongyi, S. Shiquan, S. Chuanqing, and W. Xiangkun. 2002. Studies on sampling schemes for the establishment of core collection of rice landraces in Yunnan, China. Genet. Resour. Crop Evol. 49:67–74.